ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 20-235

Matching Methods for Clustered Observational Studies in Education

Luke Keele
University of Pennsylvania

Matthew A. Lenard
Harvard University

Lindsay C. Page
University of Pittsburgh

Many interventions in education occur in settings where treatments are applied to groups. For example, a reading intervention may be implemented for all students in some schools and withheld from students in other schools. When such treatments are non-randomly allocated, outcomes across the treated and control groups may differ due to the treatment or due to baseline differences between groups. When this is the case, researchers can use statistical adjustment to make treated and control groups similar in terms of observed characteristics. Recent work in statistics has developed matching methods designed for contexts where treatments are clustered. This form of matching, known as multilevel matching, may be well suited to many education applications where treatments are assigned to schools. In this article, we provide an extensive evaluation of multilevel matching and compare it to multilevel regression modeling. We evaluate multilevel matching methods in two ways. First, we use these matching methods to recover treatment effect estimates from three clustered randomized trials using a within-study comparison design. Second, we conduct a simulation study. We find evidence that generally favors an analytic approach to statistical adjustment that combines multilevel matching with regression adjustment. We conclude with an empirical application.

VERSION: May 2020

Suggested citation: Keele, Luke, Matthew A. Lenard, and Lindsay C. Page. (2020). Matching Methods for Clustered Observational Studies in Education. (EdWorkingPaper: 20-235). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/r5hw-g721

Matching Methods for Clustered Observational Studies in Education*

Luke Keele† Matthew Lenard‡ Lindsay Page§

May 20, 2020

Abstract

Many interventions in education occur in settings where treatments are applied to groups. For example, a reading intervention may be implemented for all students in some schools and withheld from students in other schools. When such treatments are non-randomly allocated, outcomes across the treated and control groups may differ due to the treatment or due to baseline differences between groups. When this is the case, researchers can use statistical adjustment to make treated and control groups similar in terms of observed characteristics. Recent work in statistics has developed matching methods designed for contexts where treatments are clustered. This form of matching, known as multilevel matching, may be well suited to many education applications where treatments are assigned to schools. In this article, we provide an extensive evaluation of multilevel matching and compare it to multilevel regression modeling. We evaluate multilevel matching methods in two ways. First, we use these matching methods to recover treatment effect estimates from three clustered randomized trials using a within-study comparison design. Second, we conduct a simulation study. We find evidence that generally favors an analytic approach to statistical adjustment that combines multilevel matching with regression adjustment. We conclude with an empirical application.

Keywords: Causal Inference; Clustered Observational Studies; Clustered Randomized Trials; Hierarchical/Multilevel Data; Optimal Matching

*We gratefully acknowledge funding support for this work from the Spencer Foundation. We thank Wake County Public School System staff members Brad McMillan, Colleen Paeplow, and Sonya Stephens. The opinions expressed here do not necessarily reflect those of the Spencer Foundation. All errors are our own.

†University of Pennsylvania, Philadelphia, PA, Email: luke.keele@gmail.com.

‡Harvard University, Cambridge, MA, Email: mlenard@g.harvard.edu

§University of Pittsburgh, Pittsburgh, PA, Email: lpage@pitt.edu

1 Introduction

Educational interventions are often allocated to intact groups rather than to individuals. For example, a new reading program may be implemented in some schools but not in others. Such grouped treatments may be randomly or non-randomly allocated. When clustered treatments are randomly allocated, the statistical power to detect treatment effects is reduced, but the treatment effect estimates remain unbiased. Such clustered randomized trials (CRTs) have become a common research design in educational settings. Many observational studies in education share the same grouped treatment assignment mechanism as a CRT, where schools or classrooms are selected for a new program, curriculum or intervention. When clustered treatments are non-randomly allocated, the research design may be described as a clustered observational study (COS) (Page et al. 2019).

When a treatment is non-randomly allocated—grouped or not—differences in outcomes between those who are and are not treated may reflect pretreatment differences rather than the effect of the treatment itself (Cochran 1965; Rubin 1974). Such pretreatment differences may be measurable and thus constitute overt biases. Alternatively, these differences may be unmeasured and form hidden biases. In the context of any observational study, the need to remove overt bias necessitates the use of statistical adjustment methods to make treated and control groups comparable in order to estimate causal effects. In a COS, investigators often use multilevel regression models to adjust for observed confounders and remove overt bias. Regression models, however, impose strong functional form assumptions that may bias treatment effect estimates. A wide variety of matching and weighting methods have been developed as more robust alternatives to regression modeling. However, little of this development has focused on clustered observational studies.

Here, we review and evaluate a new matching method—multilevel matching—specifically designed for clustered observational studies. First, we provide a non-technical introduction to this new form

of matching. Second, we evaluate multilevel matching by attempting to recover treatment effect estimates from three CRTs. In each case, we are able to recover experimental benchmarks and show that that a combination of multilevel matching followed by regression analysis with the matched data tends to reduce bias more than either matching or regression used in isolation. Third, we conduct a simulation study based on a real clustered observational study. Our simulation results shed light on how the clustered nature of treatment assignment affects the performance of treatment effect estimation. Overall, we contribute important, new evidence on best practices for statistical adjustment in COSs. Throughout, we emphasize the practical implications of our results for applied research that utilizes the COS design.

This paper proceeds as follows. Section 2 reviews design considerations for a COS. Here we advocate for designing COSs to emulate the CRT the researcher would have designed, were it possible. Section 3 discusses the mechanics of applying multilevel matching procedures to COSs. Section 4 summarizes how to conduct outcome analyses after applying the multilevel matching approach. Section 5 reports experimental benchmarks derived from multilevel matching using data from existing CRTs. Section 6 reports on the bias-reduction potential of multilevel matching through the use of a simulation study. Section 7 presents an empirical application in the context of a COS. Section 8 concludes.

2 Clustered Observational Studies: A Review

We are focused on a form of matching that is tailored to a specific research design, to which we refer as a clustered observational study. A COS is the observational counterpart to the clustered randomized trial. That is, treatment assignment occurs at the group level, while analytical interest is focused on unit outcomes. Throughout we will use the term cluster to refer to groups for which treatment has been assigned and unit to refer to the units within clusters. Throughout, because we are focused on COSs in educational settings, we will use the terms "cluster," "group," and "school" interchangeably. We will do the same for the terms "unit" and "student." Given the

tight connection between the study design and the matching method, we first review the study design. We refer the interested reader to Page et al. (2019) for a more detailed treatment of this topic.

2.1 Target Trials and Notation

Here, we follow an approach to causal inference that advocates target trial emulation (Hernán and Robins 2016). That is, we apply design principles from randomized trials to observational studies, specifically by tying our analysis to a trial that the observational study seeks to emulate. The goal is to improve observational studies by using trial design principles even if a trial of that form is not available or feasible. Our notation is structured by the target trial of interest. As noted, for a COS, the target trial is a CRT. In the educational research context, clustered treatment assignment happens commonly due to an intervention being administered to clusters. Since we are focused on matching, we specifically seek to emulate a matched-pair CRT: a trial where schools are paired based on a similar pre-treatment characteristic (or characteristics) and then randomized to treatment and control conditions. Here, we consider two different forms that the target trial may take and how the relevant form affects the matching process.

We refer to the first target trial as Design 1. Design 1 describes a COS where complete clusters are selected for treatment, and all units within a given cluster either receive or do not receive treatment. Under Design 1, given that the assignment mechanism is at the cluster level and assignment is made on the basis of cluster-level characteristics, we expect that it will be critical to account for cluster-level covariates. Therefore, while aggregated student-level covariates could be important, student-level covariates may not contribute to the decision to assign treatment to a specific school. See Pimentel et al. (2017) for an example of this target trial.

Under Design 2 the target CRT is one where treatment is assigned to clusters but that treatment is applied to only a subset of the units within each treated cluster. Design 2 targets a CRT with student-level selection into the treatment within the schools selected for the experimental

intervention. How might a CRT of this type arise? Imagine an intervention for academically gifted students. The intervention could be assigned to entire schools but then applied only to the gifted student population within these schools. Under Design 2, student-level covariates will play a more critical role, since whether a student is exposed to treatment in a treated school will depend on student characteristics. This implies that we must adjust for covariates and match at both the school and the student levels. Hereafter, we do not focus extensively on this design, since we are unable to emulate this design in our main analysis, which is focused on experimental benchmarks.

Based on these two target trials, we use the following notation to describe a paired COS. There are S matched pairs of clusters, $s = 1, \ldots, S$, with two schools, $j = 1, 2$, one treated and one control, for $2S$ total clusters. Each school sj contains $n_{sj} > 1$ students, $i = 1, \ldots, n_{sj}$. We have a matrix of observed, pretreatment covariates, $\mathbf{x}_{sji}$, that typically describes both students and schools. That is, $\mathbf{x}_{sji}$ typically contains baseline measures such as student sex and race/ethnicity, as well as the percentage of students in the school who are proficient in reading based on test scores. School-level measures take the same value for all students in the same school. In a COS, treatment assignment occurs at the group level, which in educational contexts often means that treatment is assigned at the school level. For the j^{th} school in pair s that receives the treatment, we write $Z_{sj} = 1$, whereas if the school receives the control, write $Z_{sj} = 0$. We define causal effects using the potential outcomes framework (Neyman 1923; Rubin 1974). Prior to treatment, each student has two potential responses: (y_{Tsji}, y_{Csji}), where y_{Tsji} is observed from the ith subject in pair s under $Z_{sj} = 1$, and y_{Csji} is observed for this subject under $Z_{sj} = 0$. We only observe responses, which are a function of potential outcomes and treatment assignment: $Y_{sji} = Z_{sj} y_{Tsji} + (1 - Z_{sj}) y_{Csji}$.

2.2 Causal Identification Assumptions

In a COS, one reasonable estimand is the following student-level contrast: $y_{Tsji} - y_{Csji}$ which is the change in student-level outcomes caused by group-level treatment assignment. Many

investigators focus on the average causal effect of the form $E[y_{Tsji} - y_{Csji}]$ or the average causal effect for the treated: $E[y_{Tsji} - y_{Csji}|Z_{sj} = 1]$. These contrasts are referred to as estimands, since they refer to counterfactual quantities. Due to the fact that causal estimands are based on counterfactual quantities, we require a set of assumptions that allow us to identify these terms using observed data. Next, we outline the assumptions that are typically invoked for a clustered observational study to yield consistent estimates of causal effects.

First, our notation implicitly assumes that the *Stable Unit Treatment Value Assumption* (SUTVA) holds (Rubin 1986). SUTVA has two parts: 1) the treatment levels of Z_{sj} (1 and 0) adequately represent all versions of the treatment, and 2) a student's outcomes are not affected by other students' exposures. Under the first component of SUTVA, we must assume that while there may be some variation in how students receive the treatment, this variation in treatment corresponds to the same set of potential outcomes. Next, clustered treatment assignment is generally assumed to bolster the second component of SUTVA. For student outcomes to be affected by other students' treatment status, treatment spillovers must occur across schools. That is, only interference across schools would be problematic. While possible, in most education research settings it is unlikely that cross-school spillovers will occur. Importantly, the potential outcomes notation we outlined above allows for arbitrary patterns of interference among students within the same school, given treatment selection at the school level.

The key assumption in a COS is that treatment assignment depends on observed covariates only. Next, we provide a formal definition of this assumption. To do so, we introduce some additional notation. Specifically, we write u_{sji}, which is an unobserved, binary confounder. Formally, we must assume that:

$$\pi_{sj} = Pr(Z_{sj} = 1|\mathbf{y}_{Tsj}, \mathbf{y}_{Csj}, \mathbf{x}_{sj}, \mathbf{u}_{sj}) = Pr(Z_{sj} = 1|\mathbf{x}_{sj}).$$

In this equation, we see that the probability of being treated (π_{sj}) can depend on potential outcomes, observed data, and unobservables. However, in order to draw causal inferences we

must assume that this probability only depends on observed data. As such, the investigator asserts that there are no unobservable differences between the treated and control groups. This assumption goes by many different names, but it is often called "selection on observables," since the analyst asserts that there is some set of covariates such that treatment assignment is random conditional on these covariates (Barnow et al. 1980). Critically, this assumption is nonrefutable, since it cannot be tested with observed data (Manski 2007). If this assumption holds, potential outcomes will be independent of treatment assignment, and the causal effect of the treatment will be identified. Outside of randomization, many studies in education must invoke the selection on observables assumption in some form. Even in studies that are described as "natural experiments," where treatment is haphazardly allocated, if the investigator feels the need to control for confounders—as opposed to including covariates to improve precision—then the selection on observables assumption is in effect.

Next, we also assume that all schools have some probability of being treated such that $0 < \pi_{sj} < 1$. Under the principle of target trial emulation, this implies that a COS should have inclusion criteria like a CRT. In a CRT, a study population is defined, and randomization is then applied to this population. For example, in a CRT, the inclusion criteria might be elementary schools that have a Title I designation and that serve a student body in which over 50% of the students participate in the free lunch program. Schools that do not fall within these inclusion criteria have zero probability of being treated. In a COS, schools that have no probability of being treated or are almost certain to be treated should be removed just as they would be removed from a CRT due to the study inclusion criteria. In a COS, not doing so may manifest itself in large pre-treatment covariate imbalances which may occur due to treated clusters that are very dissimilar from any control clusters. Under such circumstances, the treated and control clusters are said to lack common support, since the covariate distributions do not overlap. The matching method we outline below allows investigators to trim data, typically treated clusters, to enforce common support overlap on covariate distributions and improve balance. Trimming treated clusters is not without consequence, however. Trimming treated clusters changes the estimand such that it only

applies to the population of clusters for which the effect of treatment is marginal: clusters that may or may not receive the treatment. Changing the estimand in this way may be unproblematic if the data do not represent a well-defined population (Rosenbaum 2012). We explore the issue of overlap through an empirical application below.

3 Multilevel Matching for Clustered Observational Studies

In a COS, the key identifying assumption is that treatment assignment depends on observed covariates only. In most cases, investigators will find that the treated and control groups differ with respect to baseline covariates, and they will need to make these groups more comparable by using a method of statistical adjustment to remove this overt bias. In this section, we describe a new form of matching—multilevel matching—designed specifically for COSs. One reasonable question to ask is why should analysts apply multilevel matching rather than other forms of statistical adjustment, such as regression? The answer is that matching tends to be more robust—especially when treated and control covariate distributions do not have good overlap (Imbens 2015). COSs in education often have relatively small numbers of treated and control schools which may exacerbate this problem. One potential solution might be to use propensity score methods such as inverse probability weighting. In practice, however, we have found that propensity score models fitted to hierarchical education data often suffer from convergence problems. When this happens, little progress can be made, since failure of model convergence means that one is not able to estimate propensity scores.

The matching method we developed does not require any model fitting, and thus is highly robust. In addition, multilevel matching allows for both covariate prioritization and sample trimming, which provide the analyst with control and flexibility to increase treated-control comparability on covariates deemed to be of critical importance. For example, an investigator may choose to more closely balance past test scores relative to other covariates such as school size. Moreover,

trimming allows the investigator to find the set of clusters and units with the highest levels of comparability. We provide more details on these aspects of the matching process below.

3.1 Matching and Optimality

Next, we turn to the mechanics of matching for a COS using data that have a multilevel structure. That is, we observe covariates for both individual students and the clusters in which students are nested. How should we match in this context? And more specifically, are there aspects of the match that need to be tailored to the specifics of a COS? We begin by discussing how standard matching methods could be applied to a COS and the drawbacks to using such standard methods.

One could apply standard matching methods to a COS in two ways. First, one could simply match at the student level and ignore the multilevel structure of the data. However, this process would match treated students in the same cluster to control students dispersed across multiple clusters. A match of this type is inconsistent with the target trial outline above. In general, standard matching methods are not designed to account for the multilevel structure of the data that is a key component of a COS. Alternatively, one might choose to pair schools and then match students within each matched school pair. The drawback to this approach is that it is not "optimal" (Keele and Zubizarreta 2017). The difficulty with using a non-optimal matching method is that balance can be improved with the data at hand by choosing better matches. Rosenbaum (1989) introduced the concept of optimality to matching. It is an important criterion for judging matching methods. Given its importance, here we provide a conceptual introduction to optimality in matching. Note that in this example, we assume individual-level treatment assignment and matching to ease exposition.

Assume there are T treated subjects, $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$, and C controls, $\mathcal{C} = \{\gamma_1, \ldots, \gamma_C\}$, with $C \geq T$. Each subject has a set $\mathbf{x}$ of observed covariates. Using a distance metric, such as the Mahalanobis distance, we calculate a distance $\delta_{\tau_t,\gamma_c} \geq 0$ between the covariates for each

treated unit τ_t and each control unit γ_c, where $t = 1, \ldots, T$ and $c = 1, \ldots, C$. These distances are recorded in a $T \times C$ distance matrix we denote δ. Table 1 contains a distance matrix from an example in Rosenbaum (2012). An *optimal* matching algorithm selects a mapping (α) to minimize the sum of the distances between each treated and control unit (Rosenbaum 1989). That is, an optimal match is one that chooses α to minimize $\sum_{t=1}^{T} \delta_{\tau_t, \alpha(\tau_t)}$. Matching algorithms that we might characterize as "greedy" or "short-sighted" lack optimality, since they may fail to minimize this total distance.

Table 1: A 5×6 distance matrix and an optimal assignment

	Control					
Treated	1	2	3	4	5	6
1	156	515	380	225	**84**	209
2	85	297	**185**	66	172	77
3	110	469	354	**143**	83	119
4	**144**	518	401	214	100	228
5	198	557	430	239	124	**210**

Based on the distances in Table 1, one logical way to match would be to first select the best match for the first treated unit, without consideration for future matches. This process is then repeated for each treated unit. This type of match would start by selecting $\delta_{\tau_1, \gamma_5} = 84$ as the first match, followed by the next match for the next treated unit, $\delta_{\tau_2, \gamma_4} = 66$, and so on, resulting in a total distance of $84 + 66 + 110 + 228 + 430 = 918$. This large covariate distance is mostly driven by the final matched pair selected: $\delta_{\tau_5, \gamma_3} = 430$. A variant of this general approach would be to randomly pick a treated unit for the first match and proceed from that randomly selected treated unit.

We can improve the overall quality of the match (e.g., the overall distance) by changing early matches to improve the quality of later matches. The matches highlighted in bold are from an optimal match which has a total distance of $84 + 185 + 143 + 144 + 210 = 766$. This better result is a function of picking earlier matches that are somewhat worse in order to improve the quality of the match overall. In short, an optimal match minimizes the covariate distances such

that *no other* match will produce a lower total distance. The key insight of this simple example is that one can improve the quality of the match by changing the type of match and the associated matching procedure.

When matching for a COS, pairing schools or using student level covariate information, as outlined above, would seem a reasonable option. However, a match of this type is not optimal (Keele and Zubizarreta 2017). Pimentel et al. (2017) provide additional simulation evidence on this point. The general problem is that it is not clear *a priori* which school-level match will produce the best student-level match, thus rendering a given school match potentially suboptimal. Next, we outline a matching algorithm that is optimal for the multilevel data structure that is a hallmark of a COS (Keele and Zubizarreta 2017; Pimentel et al. 2017).

3.2 Student Matches in Multilevel Matching

Next, we describe multilevel matching for COSs, with a focus on building intuition for the process. We refer the interested reader to Keele and Zubizarreta (2017) for a formal description of multilevel matching. Multilevel matching proceeds in a counter-intuitive way in that it explores student matches before school matches. This is done to ensure that the match is optimal. If we match on schools and do not account for student-level information, the match is ignoring critical information and thus won't be optimal.

Multilevel matching for a COS proceeds via the following steps. First, using covariates measured at the student level, student-level matches are conducted for *all* possible pairings of treated and controlled schools. For example, suppose there are 3 treated schools and 5 control schools indexed by $k_t \in \mathcal{K}_t = \{1, ..., 3\}$ and $k_c \in \mathcal{K}_c = \{1, ..., 5\}$, respectively. We first evaluate the student-pair matches across all the possible pairs of treated and control schools. Since there are 3 treated schools and 5 control schools, we conduct $3 \times 5 = 15$ student-level matches. In general, where there are N_1 treated schools and N_2 control schools, the number of such possible pairings will be $N_1 \times N_2$. Although this involves assessing a large number of possible matches, the form of these

matches is straightforward, since we are simply conducting standard student-to-student matches within each potential school pair. For these student-to-student matches we apply an optimal match based on a robust form of the Mahalanobis distance.

Next, we score the quality of each of these student matches. For each student-level match, we record a score for the pair-matched students. We denote this score as m_{k_t,k_c}. The student-level matches are scored according to two criteria. First, the match is scored based on the balance it achieves on student-level covariates, with worse scores given to matches with poor balance. Second, it is scored on the sample size it produces, with worse scores given to matches yielding smaller sample sizes. The matched sample size is measured using the harmonic mean of the treated and control group sizes. These two measures are combined additively to form the scores, m_{k_t,k_c}, which are inverted such that the best matches receive low scores, and the poorest ones receive high scores according to these criteria. These scores are then stored in an N_1 by N_2 matrix. This matrix will serve as the distance matrix for the school-level matches.

One critical point to understand about multilevel matching is that student-level covariates are incorporated into the match in an unconventional fashion. Given the multilevel structure of the data, student-level covariates can be included as either student-level measures or as aggregate measures. For example, a covariate such as student sex can be used as a binary, student-level measure or as the proportion of female students in the school. The first stage of a multilevel match avoids the need to create aggregate measures from student-level covariates. All student-level covariates will inform the score, and this score is used directly in the school-level match in the next step. However, one can also include aggregate covariates during the school matching process.

3.3 Refined Covariate Balance

At this point in the process, matched school pairs can be created using the distance matrix created from m_{k_t,k_c}, however, school-level covariates have not yet been taken into account. To incorporate

school-level covariates, refined covariate balancing is used. We now outline this process and refer the interested reader to a technical treatment in (Pimentel et al. 2015). Refined covariate balance is based on the concept of fine balance. Therefore, we begin with an introduction to fine balance.

The goal in matching is to make treated and control group distributions similar on baseline covariates. When the treated and control distributions are similar, they are often said to be "balanced." One way to balance treated and control groups is to use fine balance. A match with fine balance on a covariate is one with identical marginal distributions for this covariate. For example, in a matched set of schools, suppose we wish to fine balance a binary indicator for Title I status. Under fine balance, even if a Title I school is matched to a school that is not a Title I school in certain pairs, the match is fine balanced if the total number of Title I and non-Title I schools in the matched controls is equal to the total in the treated group. Rosenbaum et al. (2007) use the term "fine balance" when a variable has been balanced in terms of the marginal distribution even if units are not exactly matched on the variable. In terms of removing overt bias, the effect is the same since balance is a property of groups. Randomization, for example, balances the treated and control groups, on average, but says nothing about the characteristics of any single unit in those groups.

Alternatively, one can use near-fine balance. Under near-fine balance, a covariate is finely balanced when it is feasible, and otherwise the deviation from fine balance is minimized (Yang et al. 2012). For example, a match with near-fine balance on Title I status would seek to make the number of Title I schools in the control group as close as possible to the count of Title I schools in the treated group. Fine and near-fine balance were originally designed to be applied to a small number of nominal covariates. In its original form, near-fine balance was infeasible for large numbers of covariates. Refined covariate balance is a generalization of near-fine balance for many covariates. Refined covariate balance works on many variables by applying near-fine balance to an interaction of a set of nominal covariates. However, as the number of variables in the interaction increases

the number of categories explodes, and it tends to be difficult to balance every category of the interaction. Since there typically is not any way to balance every category of the interaction, the balancing occurs in an order of priority set by the analyst. Refined covariate balance proceeds in the following way.

Suppose we have three nominal covariates, arranged in descending order of balance importance. A match satisfies refined covariate balance if:

1. The match satisfies near-fine balance for the first-priority covariate.
2. The treated and matched-control distributions of the second covariate are as close as possible among matches satisfying (1).
3. The treated and matched-control distributions of the third covariate are as close as possible among matches satisfying (1) and (2).

Intuitively, refined covariate balance asks for balance in priority order, requiring near-fine balance on the most important covariate and asking for the closest possible balance on the remaining covariates, in decreasing order of importance.

An important detail to note is that refined covariate balancing only works with discrete variables. As such, school level covariates must be transformed into discrete measures. For example, we cannot match directly on a covariate that is the percentage of students proficient in reading. Instead, we would match on a coarsened version of this measure. The refined covariate balancing works best with broad categories, so the simplest approach is to split all school-level covariates at the mean or median. Assume we have 3 school level covariates: an indicator for schools where the percentage of students proficient in reading is above average, Title I status, and an indicator for a school where the percentage of students enrolled in the free lunch program is above average. Refined covariate balance seeks to near-fine balance the interaction of these three variables. The matching algorithm would do this by trying to create a matched set of schools with nearly identical numbers of schools with these features. If we designated Title I status as the highest

priority covariate, the match would seek to near-fine balance it first, and then seek to balance the interaction of the other two covariates.

3.4 The School Matches

As we noted above, for the school-level match, we use a distance matrix that is built from the results of the student-level match. In the school matching step, we find the school pairs that minimize the total sum of school-pair distances on this distance matrix, subject to a refined covariance balance constraint. That is, school matched pairs are created so that the student-level distances are minimized for the set of treated and control schools that meet the refined covariate balance prioritization established by the user. For example, if we have prioritized the indicator for above average test scores, the matching algorithm will minimize the distance on the student-level matrix subject to the constraint that the number of schools with above average tests scores is the same across the treated and control groups. If it isn't possible to achieve the exact same number of high test score schools in the treated and control groups, the algorithm will seek to make the discrepancy in the number of schools as small as possible. See Rosenbaum (2010, Ch. 10) for a nontechnical introduction to matching with constraints of this form.

Why does the multilevel matching rely on refined covariate balance methods? First, the final match needs to simultaneously use student- and school-level information. Refined covariate balance is one way to easily add school information to the result derived from the student-level match (Pimentel et al. 2017). Second, refined covariate balancing allows the user to include balance prioritization, which we consider a critical step. We discuss balance prioritization in more detail below. Finally, refined covariate balancing allows for faster computing times in a multilevel match based on integer programming (Zubizarreta and Keele 2016; Pimentel et al. 2017). Once the schools are matched, if the investigator is using Design 2, the information from the first stage of the match is used to return matched student pairs (Pimentel et al. 2017). If Design 1 is the target trial, then only the schools are matched and all treated and control students are retained.

After matching, the most common form of balance checking is estimating mean differences between the treated and control groups. This can still be done after refined covariate balance, but it is useful to remember that the match is seeking to balance the number of schools with specific covariate profiles across the groups—not the means. In the example above, we are seeking to balance the number of schools with above average test scores. As such, a test for the difference in means will not directly measure how well the match performed at this task. However, even though the algorithm doesn't directly aim to balance the means of school covariates, standardized mean differences remain a useful heuristic to judge the balance of the school covariates after matching. In addition, one can tabulate school covariate categories across the treated and control groups as an additional balance check.

3.5 Match Options

Thus far we have described the basic mechanics of a multilevel match for a COS. Next, we review aspects of the matching process that require user input. All of the choices we discuss below are optional, but they allow the investigator to override the algorithm's default settings and further fine-tune the match.

First, we consider covariate balance prioritization. Refined covariate balance allows investigators to prioritize balance on school-level covariates. Why might an investigator want to do this? If scientific knowledge dictates that certain covariates should have higher priority, balance on those covariates can be targeted for improvement. What kind of scientific knowledge do we mean? Typically, this would be knowledge of the treatment assignment process. If district officials report that certain covariates were used to select schools for the intervention, we want to ensure that treated and control schools are comparable on these variables. Covariate prioritization can be applied to individual covariates or to covariate sets. Prioritization of covariates won't improve balance overall, but it does ensure that the covariates that are deemed to be of overriding importance are better balanced. This implies that selecting covariates for prioritization is a process of managing potential tradeoffs between improving balance for important covariates that

may cause increased imbalance in other, less important covariates.

Often investigators find that, despite matching, treated and control units are not sufficiently comparable. This manifests itself as post-match balance statistics that show treated and control units as too dissimilar. In our experience, this is common in education settings—especially when the pool of control schools is fairly small. What can be done in this situation? One strategy is to improve the match via optimal subsetting in conjunction with a school-level caliper. A caliper forbids matches between units that are not sufficiently close in terms of covariate distance. Under optimal subsetting, we subset the treated group by dropping the treated schools with the worst imbalances. Subsetting is applied by using the $\underline{n}$ parameter, which specifies a minimum number of treated units that must be included in the match. For example, if there are 20 treated schools and $\underline{n}$ is set to 19, the algorithm will discard the one treated school that improves balance best among the remaining treated schools. That is, the school that contributes the most to the imbalance will be discarded. Using this parameter, analysts can drop treated schools until balance improves. In general, we recommend dropping schools one-by-one until balance is deemed acceptable.

If the investigator decides to trim one or more schools to improve balance, they need to be aware of the fact that this changes the interpretation of the estimated treatment effect. The estimated treatment effect is no longer the effect among the treated, but the treatment effect for a subset of the treated. Thus, some amount of generalizability is lost. That is, we can no longer describe the estimated effect as the treatment effect for the entire treated population. Instead, it is the treatment effect for the subsample of the treated included in the match. When this occurs, it is critical that the analyst provide descriptive statistics for both the treated units retained in the match and those discarded by the matching. This provides important information on the population for which the treatment effect is relevant.

Next, trimming treated students can occur in a different way when the investigator pairs both schools and students under Design 2. In a match of this type, some student-level trimming is almost impossible to avoid even if subsetting has not been applied. That is, unless the student

sample sizes in the control schools are substantially larger than the sample sizes in all treated schools, some of the student-level matches will involve treated groups that are larger or very similar in size to their control groups. In these settings, under the student-level pair matching used in the first step, some treated students will invariably be excluded from the match. Here the trimming is not done to enforce balance or common support but is simply a byproduct of the structure of the pair match. Under a Design 2 match, the causal estimand is a school-level contrast for a set of students within the school who are at risk for the treatment. After matching under Design 2, analysts should again present descriptive statistics for both the students included in the match and those discarded from the match.

4 Approaches to Outcome Analyses

Multilevel matching—like any form of matching—is only a method of statistical adjustment. That is, it serves to make the treated and control groups comparable on observed covariates. Matching itself does not include any method for the estimation of treatment effects. As we outline below, treatment effect estimation after matching often utilizes standard statistical models. However, nonparametric methods can be applied as well (Rosenbaum 2002, Ch. 2). Next, we briefly review regression models for clustered observational studies.

4.1 Multilevel Regression

The primary alternative to methods such as matching is the regression model. In the context of clustered data in education, the regression model of choice is typically the random-effects (RE) regression model. In an RE regression model, there are error terms at both the unit and cluster level (Murnane and Willett 2010). In the context of a COS, the basic structure of the random-effects regression model takes the following form:

$$Y_{ji} = \gamma_0 + \gamma_1 Z_j + \boldsymbol{\beta} \boldsymbol{X}_{ij} + \upsilon_j + \epsilon_{ji},$$

where the model includes separate error terms at the school (υ_j) and student (ϵ_{ji}) levels. In the model, $\boldsymbol{X}_{ij}$ represents covariates measured at baseline that are assumed to be confounders. These covariates are added to the model to remove overt biases—observable differences between the treated and untreated schools and students. That is, the RE regression model, the primary alternative to multilevel matching, is a method of covariate adjustment for a COS. Imbens (2015) argues that matching is an attractive alternative to regression modeling for treatment effects, since it tends to be more robust to a variety of data configurations. As we outline below, the RE model is also the most straightforward way to estimate treatment effects after ML matching.

4.2 Matching and Regression

Once matching is complete, the analyst can estimate treatment effects via regression models. Under this analytic strategy, the outcome is regressed on the treatment indicator using the matched data set. For this step, standard regression models can be applied. Of course, investigators may instead prefer to apply the RE model. A random intercept model will account for within-school correlations in the standard error estimates (Murnane and Willett 2010; Raudenbush 1997). If a standard regression model is used, robust variance estimators that take into account intra-cluster correlations should be used. Spiess and Abadie (2019) find that if matching is done without replacement, using regression models to estimate the treatment effect will have valid standard errors if one accounts for pair clustering. As such, analysts should also include a random effect for paired school clusters or use a cluster robust variance estimate. The only difficulty is that this approach assumes that there is a sufficiently large number of clusters to make valid inferences.

Treatment effect estimation via regression models after matching is also useful for additional bias correction. That is, any covariates that are not fully balanced by the match can also be included in the regression model to reduce bias while estimating treatment effects (Abadie and Imbens 2011). Using regression models to further reduce imbalance after matching is a well-known idea (Stuart 2010; Ho et al. 2007). See Imbens and Rubin (2015, ch 18.8) for a

complete review of how regression may be used for additional bias correction after matching. In the education literature, one recommendation—in randomized trials—is to adjust for covariates where the imbalance exceeds a standardized difference of 0.05 or more (What Works Clearinghouse 2020). A similar rule could be applied to imbalances after matching.

In sum, the basic critique of regression models is that the analyst imposes a linear functional form assumption across the full covariate space, which may induce bias especially if there is a lack of overlap in the covariate distributions. However, if regression is applied after matching, it is applied to a subsample of the data that is well-balanced and in which covariate distributions overlap. Since regression is used locally in the covariate space, the corresponding results should be less sensitive to minor changes in the specification of the regression function (Imbens and Rubin 2015, p. 417).

4.3 Nonparametric Methods

One additional drawback to relying on regression methods is that inferences depend on the number of clusters. That is, the accuracy of quantities such as p-values and confidence intervals depend on having a large sample size—i.e., a large number of clusters. In many COSs, the number of clusters may be small. Analysts can avoid large sample assumptions by using randomization inference methods. Hansen et al. (2014) developed a set of randomization inference methods for COSs. Randomization inference methods are often referred to as permutation methods, since the inferences rely on permuting the data in a manner consistent with the implied randomization. Such inferences are valid for any sample size. Randomization inference also allows investigators to use rank-based statistics which are robust in the presence of heavy-tailed distributions. Finally, randomization inference methods allow for easy implementation of sensitivity analyses. See Page et al. (2019) for more details on the use of randomization inference in the context of COSs.

5 Experimental Benchmarks

Thus far, we have outlined multilevel matching as a method for statistical adjustment in a COS. Given that this method of matching is relatively new, there is limited work on how well it performs in practice. In this section, we further explore whether multilevel matching is the best analytic approach to estimate treatment effects when conducting a COS. Here, we take a two-pronged approach. First, we evaluate methods of statistical adjustment for COSs by attempting to recover benchmark treatment effect estimates from randomized experiments. Second, we present results from a simulation study.

Here, we describe the design for our analysis that is focused on recovering experimental benchmarks. This design is often referred to as a within-study comparison (WSC) design (Wong et al. 2018). The general approach in our analysis is to use data from a randomized trial, set aside the experimental control group, and replace it with a new control group from the overall population that was not included in the original randomized trial.

We then use this new observational data set for the study. Specifically, we apply a method of covariate adjustment to make the new control group comparable to the treated group on observed covariates. After statistical adjustment is complete, we estimate the treatment effect. We then observe whether the estimated treatment effect from the observational study design is indistinguishable from the original treatment effect estimate in the randomized trial. This is often referred to as a dependent arm WSC design (Wong and Steiner 2018). Perhaps the most well-known example of this study design is from Lalonde (1986). There are also examples of this design in education (Wong et al. 2017). See Wong et al. (2018) for a general overview of WSC designs.

We apply this study design to data from three different randomized interventions from the Wake County Public School System (hereafter, Wake County), the largest school system in North Carolina and 15th largest in the nation (Snyder et al. 2019, p. 119). All three of the original

studies were CRTs conducted in Wake County elementary schools. In each case, elementary schools were able to opt into the CRT. Among the schools that agreed to participate, a treatment was randomly allocated to half of the schools on the basis of pairwise matching on a prior value of the outcome variable. The intervention was then applied to either the entire school or to whole grades within the school. In our study, we replaced the control group from the CRT with schools that did not participate in the original CRT. For each constructed data set, we then estimated treatments after adjusting for observed covariates. For each study, we apply not only multilevel matching but also mixed effects regression models. This allows us to compare matching to a clear alternative. In some of the studies, we also included treatment effect estimates based on regression models with the matched data sets.

For each method of adjustment, we then estimated several different quantities. The first two quantities are $\hat{\tau}_{rct}$ which is the estimate from the original RCT, $\hat{\tau}_{obs}$ which is the estimated treatment effect after statistical adjustment, and $\hat{\tau}_{unadj}$ which is the unadjusted treatment effect estimate from the observational data. We then calculated two measures of discrepancy between the adjusted estimates and the RCT estimate (Steiner and Wong 2018). The first is a measure of standardized bias:

$$\frac{\hat{\tau}_{obs} - \hat{\tau}_{rct}}{SD_{rct=0}}$$

where and $SD_{rct=0}$ is the standard deviation from the outcome in the control group in the RCT. This is a standard measure for studies of this type (Wong et al. 2017). Thus, we report how far the estimates from observational study methods are from the RCT estimate. We also include 95% confidence intervals for this measure of standardized bias. Next, we estimated the percentage of bias reduction, which is calculated as

$$\left(1 - \left|\frac{\hat{\tau}_{obs} - \hat{\tau}_{rct}}{\hat{\tau}_{unadj} - \hat{\tau}_{rct}}\right|\right) \times 100$$

This research design has several advantages. First, it allows us to evaluate statistical adjustment

methods in a realistic setting. Second, this design may help us to identify the covariates that are critical for adjustment. This, in turn, would help to inform future data collection and evaluation efforts. Third, our work will inform the question of whether observational studies can provide plausible estimates of causal effects when treatments are assigned at the school level. Work by Cook et al. (2008) has helped investigators gain insight into the conditions that better allow for causal inference with observational data. Our study will further our understanding of this in educational contexts. Third, success is not guaranteed: we cannot ensure in advance that our algorithm will be able to recover the experimental benchmark. The primary disadvantage of our design is that if the estimates are too noisy, it may be difficult to draw any strong conclusions. We now turn to the three randomized interventions we consider for these analyses.

5.1 Achieve3000

In our first analysis, we use data from a CRT that was designed to evaluate Achieve3000, an adaptive literacy software program intended to increase student reading proficiency (Achieve3000 2011). In the CRT, 32 district elementary schools participated in the study. These schools were sorted pairwise on 2012-13 composite reading scores and randomly assigned within pairs to use Achieve3000. The schools that were selected for the treatment condition implemented Achieve3000 for all students in grades 2-5 and outcomes were measured in school years 2013-14, 2014-15, and 2015-16. For our analysis, we first estimated the experimental benchmark using the schools in the trial. For an outcome measure, we used end-of-grade reading scores from 2016. Student-level covariates included sex, race/ethnicity, English language learner status, gifted status, and prior achievement. School-level covariates included the percentage of Hispanic or African American students, students receiving free/reduced price lunch, students classified as English language learners, students proficient in reading and mathematics, novice teachers, (e.g., first-year teachers), and white teachers as well as an indicator for Title I school classification.

Next, we created a new data set that excluded the control schools and replaced them with the 60 elementary schools that chose not to participate in the randomized trial. We excluded 8 schools

that were using a reading program similar to Achieve3000. Since schools opted into the trial, and it was administered to all students in grades 2-5 in each treatment school, we determined that Design 1 was the most appropriate target trial. As such, we matched schools, but did not pair students within school pairs. We then sought to recover the experimental estimate by adjusting for observed confounders. We performed the multilevel match using the matchMulti package in R. We implemented three different matches. The first match used the algorithm defaults. The second match prioritized school-level reading scores, since the intervention was a reading intervention. The final match prioritized balance on the covariates with the worst imbalances to produce the best overall balance. We estimated treatment effects after matching by using RE regression models with students nested within schools and schools nested within matched pairs. We then estimated the treatment effect including covariates in the model to provide additional bias correction. Finally, we estimated the treatment effect using only RE regression for adjustment in the data without matching.

Table 2 contains the results from the analysis. First, we estimated an unadjusted treatment effect. The bias for the unadjusted estimate is -0.18 which indicates that the schools that opted out of the original study tended to have lower reading scores. If we adjusted for observed confounders, using a regression model, the estimated bias is -0.08 and the 95% confidence interval now includes zero. Multilevel matching using the default settings produces an estimated bias of -0.13 standard deviations. However, the three more customized matches (Matches 2–4) produce smaller amounts of bias: -0.04, -0.05, -0.04, respectively. These estimates are 50% smaller than those from the regression model. Here, additional regression adjustment does not provide additional bias correction; these estimates are essentially identical to those based on regression alone. However, the confidence intervals are wide enough that we can't distinguish among the estimates. In general, for this first study, two conclusions are warranted. First, if we adjust for observed confounders, we produce treatment effect estimates that are statistically indistinguishable from those in the randomized trial. Second, multilevel matching, when carefully applied, produces estimates that are closest to those from the trial.

Table 2: Comparison of Statistical Adjustment Methods to Experimental Benchmark from Achieve 3000 Intervention

	Bias	95% CI	Treatment Effect Estimate	% Bias Reduction
Unadjusted	-0.18	[-0.30, -0.07]	-0.13	0
Multilevel Regression Model	-0.08	[-0.19, 0.03]	-0.03	58
Match 1 - Defaults	-0.13	[-0.25, -0.02]	-0.08	27
Match 2 - Reading Score Balance	-0.04	[-0.16, 0.08]	0.01	79
Match 3 - Overall Balance	-0.05	[-0.18, 0.07]	-0.01	71
Match 4 - Reading Score Only Balance	-0.04	[-0.16, 0.08]	0.00	76
Match 1 + Reg. Adjust.	-0.10	[-0.21, 0.02]	-0.05	47
Match 2 + Reg. Adjust.	-0.09	[-0.21, 0.03]	-0.04	52
Match 3 + Reg. Adjust.	-0.07	[-0.20, 0.05]	-0.02	60
Match 4 + Reg. Adjust.	-0.08	[-0.19, 0.04]	-0.03	59

5.1.1 Multi-Tiered System of Supports (MTSS)

In the second analysis, we use a CRT that evaluated the use of Multi-Tiered System of Supports (MTSS). MTSS is an umbrella framework designed to empower teachers to make instructional and behavioral decisions based on student need. In 2015-16, the district launched a randomized evaluation of MTSS by recruiting 88 schools and randomly assigning the framework to 44 schools. At the elementary school level—our focus here—54 schools were recruited, and 28 received MTSS. The full sample was sorted according to an achievement and behavior index with MTSS randomly assigned within pairs. The control condition was business-as-usual whereby teachers utilized existing resources to meet students' needs. Treated schools were provided with MTSS coaches who provided ongoing support to schools, ongoing technical training for all school-based MTSS teams, consulting services from MTSS program advisors and staff, and various implementation resources. Again, we use end-of-grade reading scores as the outcome measure. See the appendix for the list of student- and school-level variables used for adjustment.

The analysis mirrors the Achieve3000 CRT in that we first estimated the treatment effect from the randomized trial. We then discarded the experimental controls and replaced them with schools that did not opt into the study. We again adjusted for observed confounders using the same three

forms of multilevel matching and regression models. Table 3 contains the results. Again, the unadjusted estimate differs significantly from the experimental benchmark estimate (bias: 0.138). All three forms of multilevel matching produce estimates which are statistically indistinguishable from the experimental benchmark. However, the regression estimates are noticeably closer to the experimental benchmark (bias: -0.002). Next, we estimated treatment effects by applying regression models to the matched data and adjusting for any covariates with residual imbalances. These estimates are nearly identical to those from regression adjustment alone. What explains this result? Matching is focused on modeling the treatment assignment, while regression is a model of the outcome. Clearly some of the confounders are strongly related to the outcome (most likely prior test scores), and in this application, adjustment via a regression produces additional bias reduction.

Table 3: Comparison of Statistical Adjustment Methods to Experimental Benchmark from MTSS Intervention

	Bias	95% CI	Treatment Effect Estimate	% Bias Reduction
Unadjusted	0.138	[0.058, 0.218]	0.812	0
Multilevel Regression Model	-0.002	[-0.082, 0.078]	-0.068	99
Match 1 - Defaults	0.055	[-0.029, 0.139]	0.289	60
Match 2 - Reading Score Balance	0.036	[-0.059, 0.130]	0.167	74
Match 3 - Overall Balance	0.038	[-0.046, 0.123]	0.185	72
Match 4 - Reading Score Only Balance	0.018	[-0.071, 0.107]	0.057	87
Match 1 + Reg. Adjust.	-0.003	[-0.087, 0.081]	-0.076	98
Match 2 + Reg. Adjust.	-0.002	[-0.096, 0.092]	-0.068	99
Match 3 + Reg. Adjust.	0.002	[-0.082, 0.086]	-0.046	99
Match 4 + Reg. Adjust.	-0.004	[-0.093, 0.085]	-0.082	97

5.1.2 Nurturing for a Bright Tomorrow (NBT)

In our final benchmarking analysis, we use a randomized evaluation of Nurturing for a Bright Tomorrow (NBT). NBT was designed for use in early elementary school classrooms to provide teachers with a framework to differentiate instruction, teach advanced vocabulary and speaking skills, and build sustainable problem-solving traits. The primary goal in the study was to under-

stand whether the NBT intervention would increase enrollment in the district's gifted and talented program. District staff recruited 32 schools with below average gifted enrollments to participate in the CRT. The CRT design was again a matched-pair design in which schools that were ranked on prior gifted program enrollment rates were sorted into pairs, and NBT was randomly assigned to one school within each matched pair. NBT was applied to schools for each new kindergarten cohort. The primary outcome was a binary measure for whether the student enrolled in the gifted program in fall of the third grade year.

Our study design was consistent with Achieve3000 and MTSS. That is, we again replaced the control schools with schools that did not participate in the CRT. We matched treated schools to the new control cohort. Table 4 displays the results. Again, the unadjusted estimate differs significantly from the experimental estimate. Here, matching and regression adjustment all produce bias estimates of similar magnitudes. That is, bias under regression alone is 0.018 and bias under one of the three matches is 0.013. However, we find there is additional bias reduction when we apply regression modeling to the matched data and include covariates. Under this approach, the bias for all three methods of matching is quite low: 0.009 to -0.008.

Table 4: Comparison of Statistical Adjustment Methods to Experimental Benchmark from the NBT Intervention

	Bias	95% CI	Treatment Effect Estimate	% Bias Reduction
Unadjusted	-0.028	[-0.137, 0.082]	-0.059	0
Multilevel Regression Model	0.018	[-0.061, 0.098]	-0.013	34
Match 1: Defaults	-0.001	[-0.097, 0.096]	-0.032	98
Match 2: Reading Score Balance	-0.026	[-0.148, 0.097]	-0.057	8
Match 3: Overall Balance	-0.013	[-0.126, 0.099]	-0.045	52
Match 1 + Reg. Adjust.	-0.006	[-0.105, 0.093]	-0.038	77
Match 2 + Reg. Adjust.	-0.008	[-0.108, 0.092]	-0.039	71
Match 3 + Reg. Adjust.	0.009	[-0.078, 0.095]	-0.023	69

In sum, two broad conclusions are warranted based on these three benchmark studies. First, we were able to recover the experimental benchmark in all three cases. That is, across all three applications, we were able to use observed covariates to obtain treatment effect estimates that

are very close to the randomized trial results. Why might this be the case? While we cannot know for sure, it is our supposition that this is due to the nature of clustered observational studies. That is, selection into treatment tends to be based on school-level covariates and is often conducted by school district officials. Such scenarios typically are easier to model than applications where subjects select their own treatments based on anticipated benefits. Second, we can also draw some limited conclusions about the best approach to statistical adjustment. Consistent with much other research, we found that while matching often performs well, regression after matching can provide additional bias reduction. In the first case study, matching alone produced the best results, and additional regression adjustment was unnecessary. In the second application, regression and matching with regression produce nearly identical results. In the second and third applications, additional regression adjustment after matching was critical to improving results. However, in we must interpret the results with some caution in the first two cases, since we did not have sufficient power to distinguish between methods. Next, we conduct a simulation study to to better understand the variation in performance across matching alone, regression alone, and matching together with regression. We can design the simulation study so that matters of statistical power are not concern.

6 Simulation Study

For our simulation study, we sought to develop a design that closely mimics an actual COS in education. To that end, we based our simulation on a COS in Pimentel et al. (2017). That study investigated the effectiveness of a summer school reading intervention in Wake County, NC. More specifically, the authors evaluated the use of myON, a digital reading platform designed to increase reading comprehension among students. Schools were selected for the intervention based on a mix of factors including internet bandwidth, computer access, and geographic distribution. Given that the intervention was assigned to entire elementary schools, the design follows the COS template. Next, we outline the simulation design and how we used the myON data to structure the simulation.

We begin with a short review of details about the data set that influence the simulations. In the data, there are 18 treated schools with 1,367 students, and 26 control schools with 2,060 students. There are 5 student-level variables and 9 school-level variables. The student-level variables are reading and math test scores, and indicator variables for race/ethnicity and sex. The school covariates include the percentage of students receiving free/reduced price lunch, English language learners, novice teachers, and students proficient in math and reading. We also include school-level measures for staff turnover rates and student average daily attendance. We calculated R^2 values for both the outcome and for a propensity score model stratified by student and school-level covariates. The R^2 using student-level variables for the outcome is 0.59 and for treatment assignment the pseudo-R^2 is 0.003. The R^2 using school-level variables for the outcome is 0.004 and for treatment assignment the pseudo-R^2 is 0.28. As such, in this data, school-level variables are critical for treatment assignment, and student-level variables are strongly associated with the outcome. Our simulation utilizes this feature as we outline next.

We begin by describing the notation for our data generating process used in the simulation. We use $\mathbf{X}$ to represent both school- and student-level variables in the myON data, summarized above. Since $\mathbf{X}$ includes covariates at both levels, we omit subscripts. We use separate notation for three covariates related to achievement. We denote student-level reading scores with R_{ij}, student-level math scores as M_{ij}, and the percentage of students proficient in math and reading in each school with P_j. We then fit the following model:

$$Y_{ij} = \beta_0 + \beta_1 \mathbf{X} + \beta_2 R_{ij} + \beta_3 M_{ij}$$

where Y_{ij} are student-level reading scores and β_0 is the intercept. After fitting this model, we save $\hat{\beta}_0$ and $\hat{\epsilon}_{ij}$, the estimated residuals from this regression model. We use $\tilde{\tau}$ to denote the true treatment effect estimate in the simulation. We set the true treatment effect to be a third of a standard deviation of the raw outcome measure.

Next, we generated simulated data via the following steps. First, we generate potential outcomes

under control as

$$y_0 = \hat{\beta}_0 + 2.5R_{ij} + 2.5M_{ij} + 1.9P_j + v_1$$

where v_1 is a draw from a normal distribution that is mean zero with a standard deviation of 12. Next, we generated potential outcomes under treatment as

$$y_1 = y_0 + \tilde{\tau} + \delta_i$$

where δ_i is the following student-level model for treatment effect heterogeneity:

$$\delta_i = .01R_{ij} + .05M_{ij} + v_2.$$

Again v_2 is a draw from a normal distribution that is mean zero with a standard deviation of 2. This implies that there is a systematic treatment effect as well as student-level heterogeneity in the treatment effect. We also developed a model for the treatment assignment process. We stipulate that treatment assignment is a school-level process that is only a function of school academic performance:

$$S_j = 25 - .5P_j$$

We then generate the school-level treatment assignment indicator, Z_j, as draws from a Binomial distribution with central tendency d_j, where d_j is

$$d_j = \frac{\exp(S_j)}{1 + \exp(S_j)}.$$

There were 44 schools in the data set and 18 are treated. This treatment selection model produced between 15 and 25 treated schools in each iteration of the simulation. Next, we generate simulated outcomes based on the following equation:

$$\tilde{Y}_{ij} = Z_j y_1 + (1 - Z_j)y_0 + u_{ij} + \tilde{u}_j$$

In the equation for the outcomes, there are both student and school specific error terms. Above, $u_{ij} = \hat{\epsilon}_{ij} + v_3$, where v_3 is a draw from a normal distribution with mean zero and a standard deviation of 4. Next, $\tilde{u}_j$ is a school-level error term that is the mean of $\hat{\epsilon}_{ij}$ for each school.

There are several important features of our simulation that are critical to understand. First, potential outcomes under control are a function of the student-level test scores and, to a lesser extent, the overall quality of the school as measured by the percentage of proficient students. Second, for each student there is both a systematic treatment effect that is constant and an idiosyncratic component that is also a function of student-level test scores. Third, selection for treatment at the school level is only a function of school-level academic performance. Finally, the error term has both student- and school-level components, as we would expect in a COS.

In each simulation scenario, we apply three different estimation methods. First, we use multilevel matching to adjust for observed confounders. After matching, we estimate the treatment effect by applying a RE regression model to the matched data. Second, we use multilevel matching, but after matching, we estimated the treatment effect using an RE regression model that includes all the baseline covariates. Third, we estimated the treatment effect using a RE regression model alone. For the estimates from the RE regression alone, we use the complete control group.

Next, we conducted four simulation studies, where we varied the specification for the three estimators. In the first scenario, we omitted the school-level academic performance measure P_j from the ML matching and the RE regression. However, we include the school-level academic performance covariate in the RE regression model applied to the matched data. We refer to this scenario as the "Student In" case, since we condition on student-level test scores, and we omit the key school-level confounder. In the second scenario, we omitted the student-level test score measures R_{ij} and M_{ij} from the matching estimator and the RE regression model. These covariates are included in the post-match regression model. We refer to this scenario as the "School In" case, since we condition on the key school-level covariate, and we omit the key student-level confounders. In the next scenario, both student- and school-level covariates are

included, so all three methods of estimation are fully specified. In the final scenario, we omit both student- and school-level measures from the matching and RE regression estimator. However, these covariates are included in the regression model that further adjusts after matching. Thus, we seek to understand how misspecification affects estimation methods in the context of the COS. Specifically, we designed the study such that the key school-level covariate determines outcomes and treatment assignment, while the key student-level variables only determine outcomes. As such, we can understand how methods that model either assignment, outcomes, or both are affected by omitting these covariates from the specification. Note that in all the scenarios, we still include all the other baseline covariates in either the regression model or the match. However, these covariates are, by construction, not determinants of either the outcomes or treatment assignment process, except insofar as they are correlated with the key covariates.

For each scenario, we repeated the simulation 1,000 times, and we report the bias for each method. We measured bias as absolute standardized bias. That is, we calculate the absolute value of the average bias and divide it by the standard deviation of the control group from the original data. Finally, we comment on three additional considerations. First, the outcome model is linear—as such, the data generating process should be amenable to regression modeling. Second, simulations of this type are difficult to conduct with matching. Normally, the matching process would included iterative fine-tuning to reduce imbalances. Here, the imbalances will vary stochastically from simulation to simulation, and we cannot practically adjust the matching parameters to match the variation in the imbalances over a thousand simulations. In theory, matching would perform better if we were able to adjust the match in each case to improve balance. Finally, standard practice is to include only imbalanced regressors in the regression model after matching. Here, we include all variables, since again balance checking is impractical in a simulation.

6.1 Simulation Results

Table 5 contains the results from the four simulation scenarios. First, we review the results from the Student In scenario. Here, we observe that if we do not make any adjustment for covariates, the bias is on average nearly 0.30 standard deviations—a rather large bias that would completely mask the true treatment effect. We find that the RE regression model is robust to misspecification in this scenario. That is, so long as the key outcome-level covariates are included, the RE regression model displays little bias. Matching alone, however, performs poorly with an average bias of 0.13. However, post hoc regression adjustment removes this bias. Next, in the School In scenario, matching has little bias—0.03 of a standard deviation. Now the RE regression is biased—0.18 of a standard deviation. Clearly, regression performs poorly when we only condition on the key school-level covariate that strongly determines treatment assignment. When there is no mis-specification, all three methods perform similarly. Finally, when we omit both sets of variables, the RE regression model performs worst. Here, the doubly robust option performs best. However, the bias from matching alone is much smaller than that of regression alone. This is due to the fact that the school-level variables that are included are more strongly correlated with school-level academic performance than the other student-level measures. As such, adjusting for these other school-level covariates allows for some bias reduction. Overall, the results of the simulation comport with conventional wisdom. That is, when key confounders are in the outcome model, outcome modeling will perform well. When key confounders predict treatment, methods such as matching perform well. What is novel about our simulation study is that in an education context, matching appears to be more robust, since school-level predictors often serve as important proxies for each other in treatment assignment. The smaller correlations among student-level variables make them poor proxies when those key variables are missing.

Finally, these results may also help explain the variation in the results of the experimental benchmark results. Matching performed well in the first application. This suggests that the school level covariates were critical in the analysis. In the second analysis, regression performed well, which

Table 5: Bias in Treatment Effect Estimates For Three Simulation Scenarios

	Student In	School In	Both In	Both Out
Unadjusted	0.28	0.28	0.28	0.27
Regression	0.03	0.18	0.04	0.18
Multilevel Match	0.13	0.03	0.01	0.10
Multilevel Match + Regression	0.05	0.04	0.04	0.05

suggests that student-level covariates were critical. In this application, the additional bias reduction from regression after matching was key. In the third application, we needed both matching and regression for the largest amount of bias reduction. That implies there were key imbalances in school- and outcome-level covariates that needed to be accounted for.

7 Application

Finally, we present an empirical application. We use this empirical application to demonstrate one key strength of matching in the context of a COS. When treatments are assigned at the school level, this often results in smaller samples, which may exacerbate a lack of overlap in the treated and control covariate distributions. Here, we use the empirical application to demonstrate how matching can reveal overlap issues in data in a way that regression modeling cannot. We begin by reviewing the details of the application.

In 2015-16, Wake County followed many other large districts by implementing a school turnaround strategy known as the Elementary Support Model (ESM). ESM was designed to increase support for the district's chronically lowest performing elementary schools. To identify these schools, district staff developed an index that accounted for academic, human capital, behavioral, and socioeconomic indicators and was averaged over the previous three years. The 12 elementary schools ranking at the bottom of this index were non-randomly assigned to the ESM treatment condition and received a range of supports over the next three years, including governance reform, additional staffing, and instructional coaching. The goal of ESM was to help these 12 schools transition out of chronically underperforming status. A district evaluation of ESM used

an unspecified matching procedure that resulted in an analytic sample of the 12 ESM schools and 12 matched comparison schools (Paeplow et al. 2019). Here, we present a brief analysis of the ESM intervention to compare the utility of multilevel matching to an analysis based on regression modeling alone.

Our analytic sample consists of the final year of a three-year panel that spans 2015-16 to 2017-18 and includes all schools and students in the district. In the fall prior to ESM's launch, the district had 104 elementary schools—12 ESM and 92 non-ESM. Our data includes school and student-level variables merged into a single data source. Student-level covariates include student sex, an indicator for limited English proficiency, an indicator for being advanced in grade, student race, and prior end-of-year test scores in reading and mathematics. School-level covariates include the percentage of students proficient in reading, mathematics, and science; magnet status, Title 1 status, and measures for the percentages of students receiving free/reduced lunch, students classified as English language learners, African American or Hispanic students, and National Board certified teachers. The first column of Table 6 contains balance statistics for this data set before matching. First, it is immediately obvious that there are very large differences between the treated and comparison schools. The smallest standardized difference is 0.43 and the largest is near 3. In fact, for the three measures of academic proficiency, all the standardized differences are larger than 2, and two of those are nearly 3. That implies that the means of these covariates differ by more than 2 standard deviations.

If our adjustment strategy were based on regression modeling alone, we would simply adjust for these covariate differences. We would have little sense of how much improvement such regression adjustment made in terms of comparability. That is, these groups might still be incomparable. Here, we implemented a multilevel match. In the match, we prioritized the large imbalances in academic proficiency as a set, and balanced on 3 additional, strongly imbalanced covariates. The results are contained in the second column of Table 6. We find that balance is dramatically improved. For the 3 measures of academic proficiency, the standardized differences are now near

one instead of 3, and balance is generally improved. However, the rough rule of thumb is that after matching, standardized differences should fall below 0.10. In the context of ESM, only two covariates approach that threshold. Next, we implemented a match that trims the treated schools. We trimmed iteratively and stopped after dropping six treated schools. The final column of Table 6 contains the balance statistics for this match. Again, while the improvements in balance are large, we are nowhere near the standard benchmark for standardized differences. What have we learned? Here, the matching process provides clear evidence that the treated schools are quite different from the control schools. Moreover, even a subset of the treated schools remain very different from the controls. Figure 1 contains box plots for two covariates after matching. The lack of overlap in the covariates distributions is quite obvious.

What should one do? Here, the best option is to expand the control pool to include schools outside of the district. We would caution against estimating treatment effects even after matching given clear levels of incomparability. Matching readily reveals the large differences between treated and control schools and clearly warns against invalid inferences in a way that regression does not. The ability to detect such large differences is critical in the context of COS applications, since treated and control pools tend to be smaller. While this ability to detect large differences has long been a strength of matching, multilevel matching makes these tools available to applications that conform to the COS template.

8 Conclusion

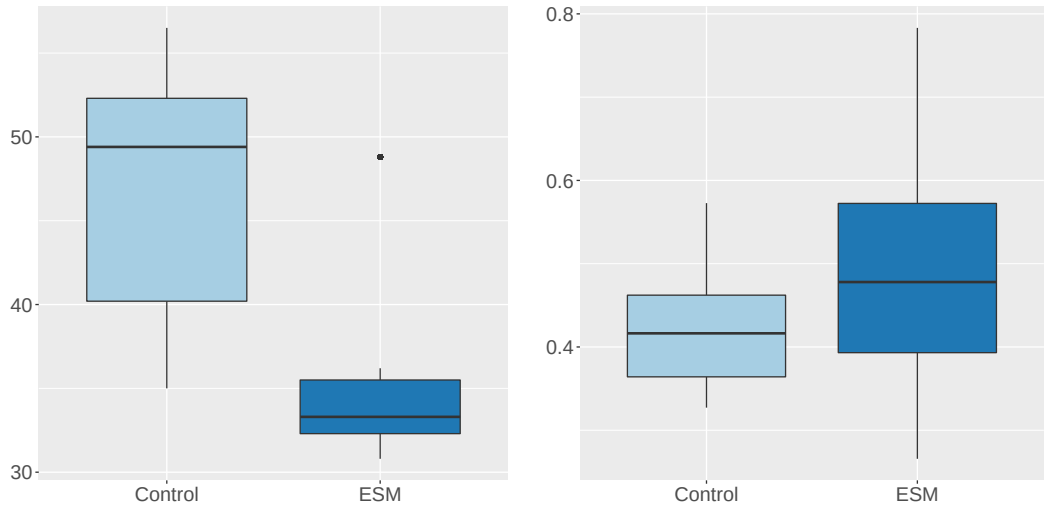
Clustered observational studies are a common study design in educational research. In a COS, treatments are non-randomly assigned to clusters such as schools or classrooms. The non-random assignment of treatment implies that investigators must use statistical methods to make treated and control groups comparable in terms of observed covariates. Traditionally, multilevel regression models have been widely used for this task, since matching methods for this type of study design are relatively new. Here, we provide readers with an introduction to multilevel matching, a new

Table 6: Balance Statistics for Elementary Support Model Schools in Wake County

	Unmatched Std. Diff.	Matched Std. Diff.	Matching + Trimming Std. Diff.
Proficient Reading	-2.79	-1.16	-0.46
Proficient Math	-2.80	-1.25	-0.61
Proficient Science	-2.37	-0.96	0.04
Magnet	0.43	-0.18	-1.07
Title I School	1.38	0.00	0.00
Student With Free Lunch	2.72	1.15	0.77
LEP	1.49	0.86	0.62
Black Students	1.95	0.37	0.62
Hispanic Students	1.51	0.81	0.31
Beginner teachers	1.06	1.10	1.52
White Teachers	-1.46	-0.79	-0.50
Black Teachers	1.42	0.75	0.36
National Board Certified	-0.88	-0.33	-0.44

form of matching designed specifically for clustered observational studies. In a multilevel match, schools are paired using both school-level covariates and summary statistics based on student-level covariates. More specifically, students are first matched across treated and control schools. The quality of these matches is then used to inform a school-level match.

Next, we performed two different evaluations of multilevel matching. In the first evaluation, we recovered experimental benchmarks from three clustered randomized trials. Next, we conducted a simulation study where we constructed the simulations based on data from a real COS application. In the first study, we found variation in results that were explained by patterns in the simulations. In the simulation, we found that regression modeling performed well when only student-level covariates were available, while multilevel matching was key when school-level covariates were available. However, across these different analyses, we found that often the most effective strategy was to combine multilevel matching with regression modeling using the matched data set. This approach tended to further reduce bias compared to matching or regression used in isolation. Finally, using a real data application, we demonstrated how matching can clearly reveal overlap in covariate distributions more readily than regression modeling. When we find that balance



(a) Math Scores (b) Proportion African American

Figure 1: Covariate Distributions After Matching

and overlap are still poor *after* matching, we should recognize that a key incomparability exists between treated and control units. Such incomparability would not be readily apparent following regression analysis alone.

Multilevel matching is not without its limitations. Multilevel matching does provide researchers with many degrees of freedom in the analysis. Researchers might be tempted to trim treated units or prioritize covariates with an eye toward treatment effect estimates rather than balance metrics. The primary way to prevent this is the use of analysis plans, where match options are specified and implemented without reference to outcomes.

References

Abadie, A. and Imbens, G. W. (2011), "Bias-corrected matching estimators for average treatment effects," *Journal of Business & Economic Statistics*, 29, 1–11.

Achieve3000 (2011), "Research to Practice: How Achieve3000 Differentiated Literacy Solutions Use Research to Prepare Students to Thrive in the 21st Century," Technical Report.

Barnow, B., Cain, G., and Goldberger, A. (1980), "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies*, eds. Stromsdorfer, E. and Farkas, G., San Francisco, CA: Sage, vol. 5, pp. 43–59.

Cochran, W. G. (1965), "The Planning of Observational Studies of Human Populations," *Journal of Royal Statistical Society, Series A*, 128, 234–265.

Cook, T. D., Shadish, W., and Wong, V. C. (2008), "Three conditions under which observational studies produce the same results as experiments," *Journal of Policy Analysis and Management*, 27, 724–750.

Hansen, B. B., Rosenbaum, P. R., and Small, D. S. (2014), "Clustered Treatment Assignments and Sensitivity to Unmeasured Biases in Observational Studies," *Journal of the American Statistical Association*, 109, 133–144.

Hernán, M. A. and Robins, J. M. (2016), "Using big data to emulate a target trial when a randomized trial is not available," *American journal of epidemiology*, 183, 758–764.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007), "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political analysis*, 15, 199–236.

Imbens, G. W. (2015), "Matching methods in practice: Three examples," *Journal of Human Resources*, 50, 373–419.

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference For Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge, UK: Cambridge University Press.

Keele, L. J. and Zubizarreta, J. (2017), "Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the Effectiveness of Private Schools Under a Large-Scale Voucher System," *Journal of the American Statistical Association*, 112, 547–560.

Lalonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604–620.

Manski, C. F. (2007), *Identification For Prediction And Decision*, Cambridge, Mass: Harvard University Press.

Murnane, R. J. and Willett, J. B. (2010), *Methods matter: Improving causal inference in educational and social science research*, Oxford University Press.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science*, 5, 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).

Paeplow, C., Singh, M., and Scrimgeour, M. (2019), "Elementary Support Model Implementation and Outcomes: 2014-15 to 2017-18," *Wake County Public School System*.

Page, L. C., Lenard, M., and Keele, L. (2019), "The Design of Clustered Observational Studies in Education," Unpublished Manuscript.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), "Large, Sparse Optimal Matching with Refined Covariate Balance in an Observational Study of the Health Outcomes Produced by New Surgeons," *Journal of the American Statistical Association*, 110, 515–527.

Pimentel, S. D., Page, L. C., Lenard, M., and Keele, L. J. (2017), "Optimal Multilevel Matching Using Network Flows: An Application to a Summer Reading Intervention," *Annals of Applied Statistics*, In press.

Raudenbush, S. W. (1997), "Statistical analysis and optimal design for cluster randomized trials." *Psychological Methods*, 2, 173.

Rosenbaum, P. R. (1989), "Optimal Matching for Observational Studies," *Journal of the American Statistical Association*, 84, 1024–1032.

— (2002), *Observational Studies*, New York, NY: Springer, 2nd ed.

— (2010), *Design of Observational Studies*, New York: Springer-Verlag.

— (2012), "Optimal Matching of an Optimally Chosen Subset in Observational Studies," *Journal of Computational and Graphical Statistics*, 21, 57–71.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007), "Mimimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatment for Ovarian Cancer," *Journal of the American Statistical Association*, 102, 75–83.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 6, 688–701.

— (1986), "Which Ifs Have Causal Answers," *Journal of the American Statistical Association*, 81, 961–962.

Snyder, T. D., de Brey, C., and Dillow, S. A. (2019), "Digest of Education Statistics 2017, NCES 2018-070." *National Center for Education Statistics*.

Spiess, J. and Abadie, A. (2019), "Robust Post-Matching Inference," Unpublished Manuscript.

Steiner, P. M. and Wong, V. C. (2018), "Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons," *Evaluation review*, 42, 214–247.

Stuart, E. A. (2010), "Matching Methods for Causal Inference: A review and a look forward," *Statistical Science*, 25, 1–21.

What Works Clearinghouse (2020), *What works clearinghouse standards handbook, version 4.1*, Washington, D.C.: Institute of Education Sciences.

Wong, V. C. and Steiner, P. M. (2018), "Designs of empirical evaluations of nonexperimental methods in field settings," *Evaluation review*, 42, 176–213.

Wong, V. C., Steiner, P. M., and Anglin, K. L. (2018), "What Can Be Learned From Empirical Evaluations of Nonexperimental Methods?" *Evaluation review*, 42, 147–175.

Wong, V. C., Valentine, J. C., and Miller-Bains, K. (2017), "Empirical performance of covariates in education observational studies," *Journal of Research on Educational Effectiveness*, 10, 207–236.

Yang, D., Small, D. S., Silber, J. H., and Rosenbaum, P. R. (2012), "Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes," *Biometrics*, 68, 628–636.

Zubizarreta, J. R. and Keele, L. (2016), "Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the Effectiveness of Private Schools Under a Large-Scale Voucher System," *Journal of the American Statistical Association*, 112, 547–560.